

Classification of protein crystallization images using Fourier descriptors

Christopher G. Walker, James Foadi and Julie Wilson*

York Structural Biology Laboratory, University of York, York YO10 5YW, UK. Correspondence e-mail: julie@ysbl.york.ac.uk

The two-dimensional Fourier transform (2D-FT) is well suited to the extraction of features to differentiate image texture, and the classification of images based on information acquired from the frequency domain provides a complementary method to approaches based within the spatial domain. The intensity, I , of the Fourier-transformed images can be modelled by an equation of power law form, $I = Ar^\alpha$, where A and α are constants and r is the radial spatial frequency. The power law is fitted over annuli, centred at zero spatial frequency, and the parameters, A and α , determined for each spatial frequency range. The variation of the fitted parameters across wedges of fixed polar angle provides a measure of directionality and the deviation from the fitted model can be exploited for classification. The classification results are combined with an existing method to classify individual objects within the crystallization drop to obtain an improved overall classification rate.

© 2007 International Union of Crystallography
Printed in Singapore – all rights reserved

1. Introduction

The number of protein sequences produced each year by genome projects considerably outnumbers those proteins for which the macromolecular structure has been solved by either X-ray crystallography or nuclear magnetic resonance. The main bottleneck for X-ray crystallography is the process of crystallization as many experiments may need to be performed in order to identify the conditions that produce diffraction-quality crystals. In structural genomics centres, robotic systems can perform tens of thousands of crystallization experiments a day and the results must be assessed repeatedly over a period of time (Mayo *et al.*, 2005). Visual inspection of the images recording these results is a monotonous and time-consuming task and there is an urgent need to increase the rate at which they are evaluated. The high speed with which images are acquired places considerable constraints upon the options available for processing the images and computationally expensive methods cannot currently be considered. However, a number of projects to automate the identification of crystals growing within protein crystallization drops have been instigated (Wilson, 2002, 2004; Saitoh *et al.*, 2005; Spraggon *et al.*, 2002; Bern *et al.*, 2004; Pan *et al.*, 2006; Cumbaa *et al.*, 2003). The identification of crystals is the primary aim, but reliable classification of other experimental outcomes would provide information for subsequent trials (Jurisica *et al.*, 2001; Page & Stevens, 2004).

In most cases the image analysis software is developed for a particular imaging system. Our software, *ALICE* (*Analysis of Images from Crystallization Experiments*), is designed to be generic software that can be adapted to the many different systems available. Currently, individual objects are identified

within the crystallization drop and used to classify a crystallization image into a number of categories (Wilson, 2002, 2004). The method does not take into account the relationship between objects or their spatial arrangement and therefore higher classification rates can be achieved by combining this approach with the use of textural information obtained from the crystallization drop as a whole.

Although automated analysis of protein crystallization images has been largely carried out within the spatial domain, Fourier analysis has been used by a number of authors in other image analysis applications (Page & Stevens, 2004; Delibasis *et al.*, 1997). Since the Fourier transform (FT) can readily identify periodic and directional structure over large regions, it has been applied widely in the classification of textures. Saitoh *et al.* (2005) and Spraggon *et al.* (2002) also use textural information, but only Bern *et al.* (2004) make use of Fourier transforms to classify crystallization images. In an alternative approach, Pan *et al.* (2006) use features based on texture and the Gabor wavelet decomposition to classify individual blocks within the crystallization image.

Various approaches to the analysis of the FT can be taken. For instance, Liu (1997) used a 'Wold decomposition algorithm' which attempts to mimic the ability of the human brain to identify different textures. However, this method is computationally expensive and is therefore unsuitable for crystallization-image classification. Gonzalez & Woods (2002) describe the most common technique for extracting parameters from the FT for use in image classification. This method involves determining the intensity of the power spectrum of the FT as a function of the polar coordinates, (r, θ) , where r is measured from the zero spatial frequency point. The parameters involved in a least-squares fit of the

functions obtained by integrating over r and θ , as well as statistical values determined from the function, provide a series of extracted features that can be used to classify the images.

2. Imaging system

ALICE is being developed in collaboration with the Oxford Protein Production Facility (OPPF) at the University of Oxford, where it is now used routinely to annotate images. Many of the images shown were supplied by the OPF where crystallization experiments are performed in 96-well Greiner plates (micro-titre format) and the images are taken using an automated Oasis 1700 imaging system (Veeco, Cambridge, UK). Native images are $1024 \times 1024 \times 8$ bit bitmap (BMP) images (~ 1 megabyte in size, corresponding to a pixel width of about $3 \mu\text{m}$).

3. Power spectra

Given an image, $f(x, y)$, we follow the approach of van der Schaaf & van Hateren (1996) and rescale according to

$$\tilde{f}(x, y) = \frac{f(x, y) - \mu}{\mu}, \quad (1)$$

where μ is the mean intensity value of the image before the Fourier transform is applied. This normalizes the image to account for varying intensities caused by changes in drop curvature and ensures that the zero frequency component does not dominate the transformed image. The power spectrum, $S(u, v)$, of the image is then calculated as

$$S(u, v) = \frac{|F(u, v)|^2}{N}, \quad (2)$$

where u and v are the spatial frequencies in the x and y directions, respectively, N is the number of pixels in the image and $F(u, v)$ is the FT of $\tilde{f}(x, y)$ in equation (1). By expressing the spectrum in polar coordinates, as $S(r, \vartheta)$, where $u = r \cos \vartheta$ and $v = r \sin \vartheta$, and summing over different angles and radii, we can obtain a more global description in terms of the functions $S(r)$ and $S(\vartheta)$, where

$$S(r) = \sum_{\vartheta=0}^{\pi} S(r, \vartheta) \quad (3)$$

and

$$S(\vartheta) = \sum_{r=R_{\min}}^{R_{\max}} S(r, \vartheta). \quad (4)$$

In studies of the power spectra of natural images, *i.e.* images of phenomena that can be observed by eye, such as countryside and buildings, van der Schaaf & van Hateren (1996) found that the intensity of the radial power spectrum tends to depend on the inverse square of the spatial frequency, r . Similarly, for most of our crystallization images, we found that spectral power and spatial frequency are related by a power law, *i.e.* $S(r) = Ar^\alpha$. Fig. 1 shows a plot of the spectral power as

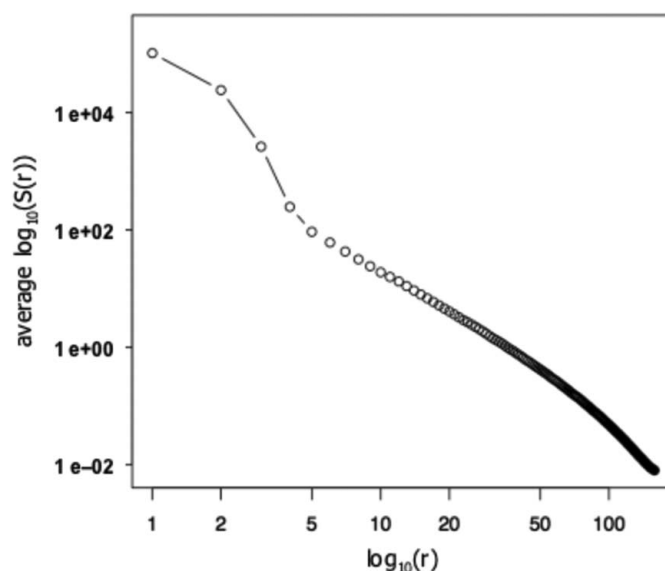


Figure 1

The average $\log_{10}[S(r)]$ over all images in the training set plotted against $\log_{10}(r)$ where, for each image, the intensity, $S(r)$, was calculated for ϑ in $[0, \pi]$ for each spatial frequency, r .

a function of spatial frequency, averaged over the 700 images in our training set. The almost straight line in the log-log plot supports the power law model for the variation of the intensity of the power spectra with radial distance.

The intensity reduces as $1/r^\alpha$, where α is typically ~ 2 , so we have transformed the power spectra in order to display both low and high spatial frequency information, and all Fourier images show $r^2 S(r, \vartheta)$.

4. Pre-processing of images

In contrast to other authors, who have studied parts of an image individually (*e.g.* 32×32 pixel blocks) in order to classify texture (Liu, 1997; Delibasis *et al.*, 1997), we have chosen to examine the texture of the whole crystallization drop for speed and simplicity. In many images there is a strong shading effect due to the curvature of the drop, as can be seen in Fig. 2. The contrast with the intensity outside the drop causes a significant contribution over a wide range of spatial frequencies and can dominate the FT. Moreover, the drop boundary may contain straight sections that mimic the presence of crystals. In order to prevent this, Bern *et al.* (2004) performed the FT over a rectangular area within the crystallization drop. However, such an approach means that a significant fraction of the drop is not considered and crystals growing at the edge of the drop (which they often do) will be missed.

The FT assumes periodic boundary conditions in both x and y directions, whereas the intensity at the left-hand edge of an image is usually different from the right (and similarly for the top and bottom). The lack of periodicity results in a cross centred at zero spatial frequency (see Fig. 2c), often the most prominent feature in the image. In order to reduce this effect,

Liu (1997) used a 'Gaussian tapering window', whereby each pixel in the image is multiplied by the function

$$g(x, y) = \exp[(x - X)^2/\sigma^2] \exp[-(y - Y)^2/\sigma^2], \quad (5)$$

where σ is related to the width of the Gaussian function centred at (X, Y) .

Fig. 2(*d*) shows the result when a Gaussian filter is centred on the crystallization drop. The contribution to the FT from both the drop boundary and the edges of the image is greatly reduced, as can be seen by comparing (*c*) and (*f*) in Fig. 2. In Fig. 2(*f*), the 'spoke' due to the crystal is now apparent (see next section). However, we found it difficult to tailor the filter to suppress strong boundaries without obscuring objects that lie close to the edge of the drop and, in fact, better results were

obtained using a mask for the drop. A rough mask for the drop is obtained quickly and this can be refined to fit the drop boundary as shown in Fig. 3 (Wilson & Berry, 2005). A close-fitting mask prevents the drop boundary dominating the FT.

A frequent problem in the analysis of protein crystallization images is the appearance of a line across the image (Wilson & Berry, 2005). This is an artifact of certain crystallization plates and can be seen in Fig. 4(*a*). The line is roughly horizontal (or vertical depending on how the plates are imaged) and therefore causes a strong vertical (or horizontal) response in the FT. Furthermore, the periodic variation in brightness across the plate can result in dots at spatial frequencies in the horizontal direction, as can be seen in Fig. 4(*b*). Thus, when analysing the FT and extracting descriptors for classification, those parts of the FT close to the horizontal and vertical axes should be disregarded.

5. Feature extraction

In order to determine different descriptors for the classification of crystallization images, the functions $S(r)$ and $S(\vartheta)$ can be calculated over different radii and angles. We found that $S(r)$ varied significantly between images at low spatial frequency, but changed relatively little at high spatial frequency due to the nature of the relationship between spectral power and spatial frequency. We therefore model $S(r)$ for two separate regions, as shown in Fig. 5. That is, we create two models, $S_1(r)$ for $r_1 < r < r_2$ and $S_2(r)$ for $r_2 < r < r_3$, where $r_2/r_1 = r_3/r_2$ so that the log of the spatial frequency changes by the same amount for each region.

The FT can be highly directional for images that contain crystals or crystal

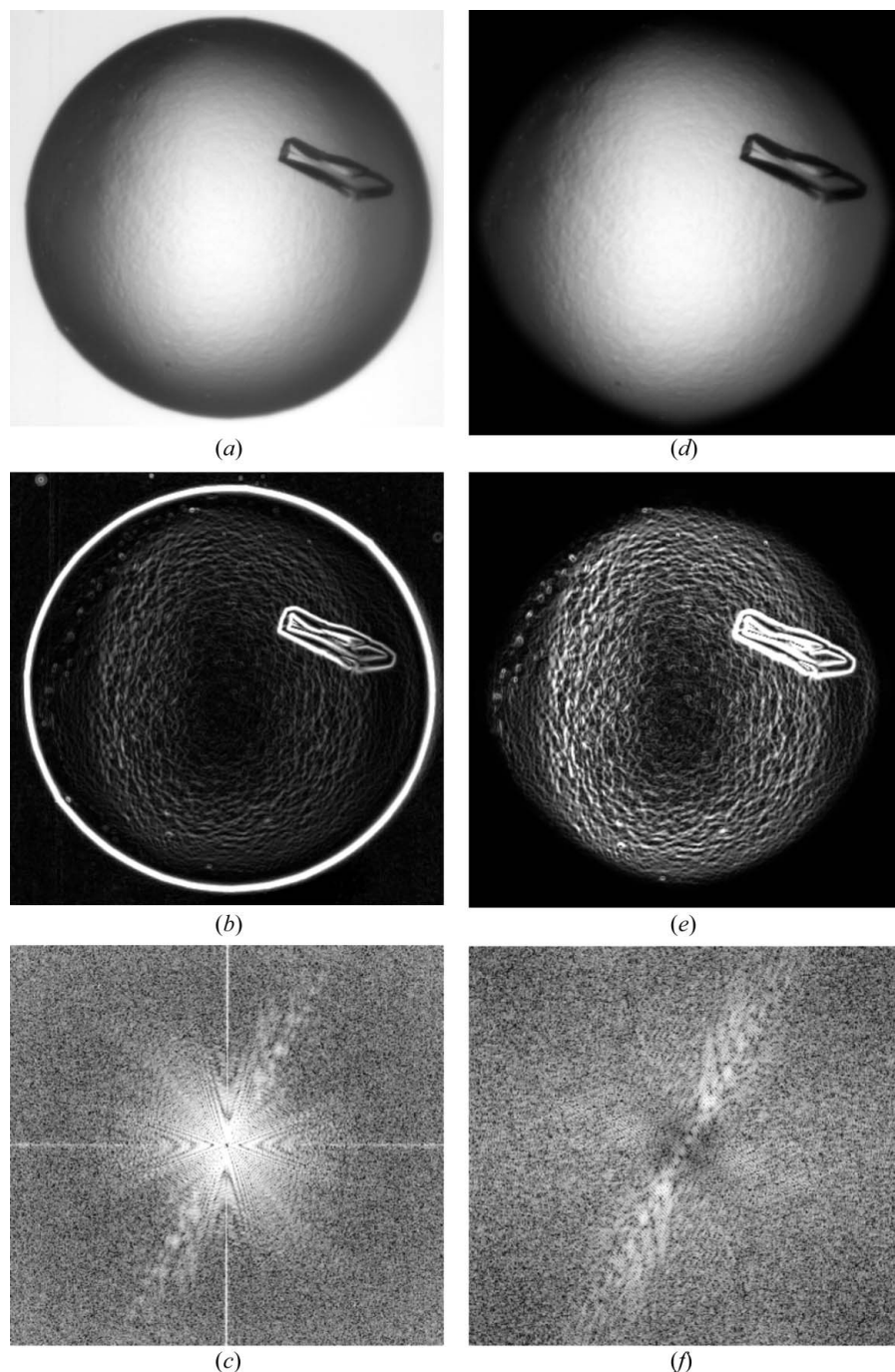
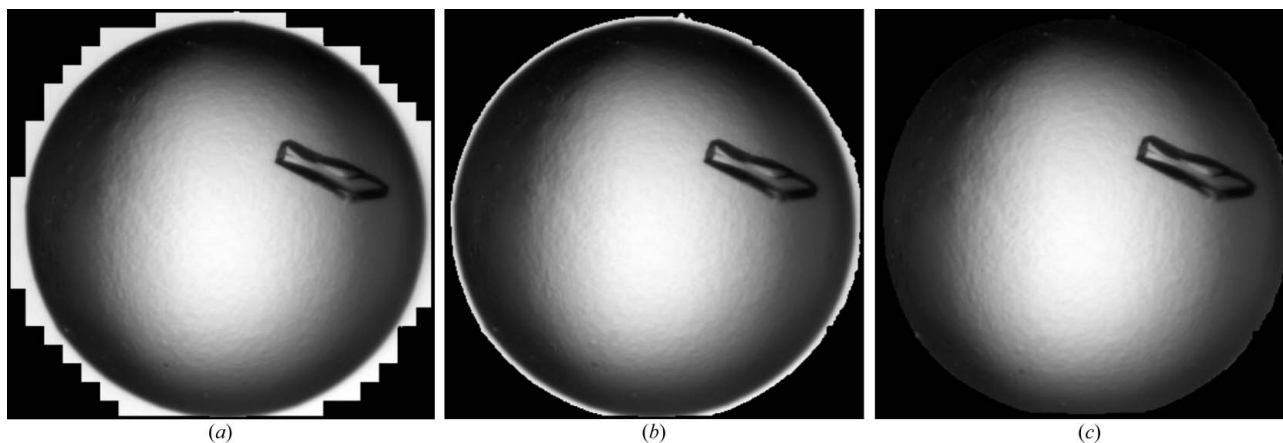
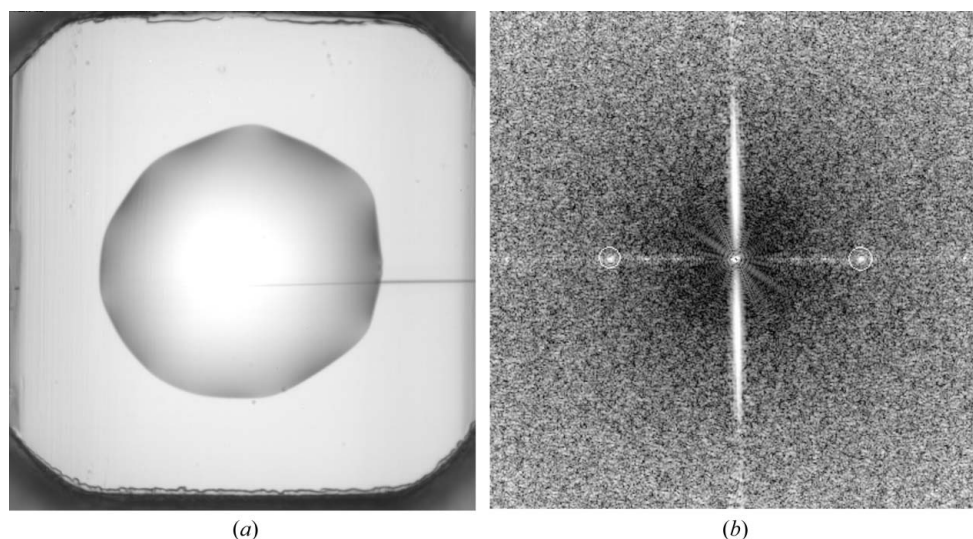


Figure 2

The Sobel gradient magnitude image obtained from the image in (*a*) is shown in (*b*). The Fourier transform of the Sobel image is shown in (*c*) and exhibits the typical cross, centred at zero spatial frequency due to non-periodicity. In (*d*) the image is shown after a Gaussian filter has been applied, with the corresponding Sobel and Fourier images in (*e*) and (*f*), respectively.

**Figure 3**

A rough mask for the drop obtained on a coarse grid is shown in (a). In (b) the mask has been expanded to the drop boundary. The bright halo occurs because the drop boundary is identified by high gradient values; expanding the mask by a few pixels removes this, as shown in (c).

**Figure 4**

The horizontal line across the image in (a) is due to the manufacturing process for the crystallization plates in which the drop sits. The strong vertical response is clear in the FT of the image shown in (b). The smaller spokes in the FT correspond to the straight sections drop boundary in this image. The dots, probably due to the periodic variation in brightness across the plate, are highlighted by circles.

clusters with 'spokes' of high intensity radiating out from the zero spatial frequency point. Fig. 6 shows a drop containing a crystal together with the FT of the image showing a definite spoke. Such spokes result in peaks of high intensity in the function $S(\vartheta)$ and the number and intensity of the peaks in the function $S(\vartheta)$ are used to obtain descriptors of directionality. Statistical measures calculated from the fitted values A and α over different polar angles provide further information for classification.

Sharp edges mean that spokes due to crystals often exhibit strong intensity oscillations along the axis of the spoke giving them the striped appearance seen in Fig. 6(b). This effect could perhaps be exploited to discriminate between spokes due to crystals and those originating from the edge of the drop where the decrease in intensity is usually more gradual due to the curvature of the drop.

5.1. Radial descriptors

The information in an image is primarily determined by contrast as a function of orientation and the total power of a spectrum can be estimated from the root-mean-square (r.m.s.) contrast, $c_{\text{r.m.s.}}$, of the input image where

$$c_{\text{r.m.s.}}^2 = \frac{1}{N} \sum_{(u,v)} |F(u,v)|^2 = \sum_{(u,v)} S(u,v). \quad (6)$$

However, van der Schaaf & van Hateren (1996) found that the variation of the measured contrast is dominated by the variation of the power at a few low-frequency coordinates and suggested alternative measures for the contrast that are less sensitive to single frequency components. They suggested the 'whitened contrast', c_w , determined by summing the power spectrum multiplied by the square of the spatial frequency, r ,

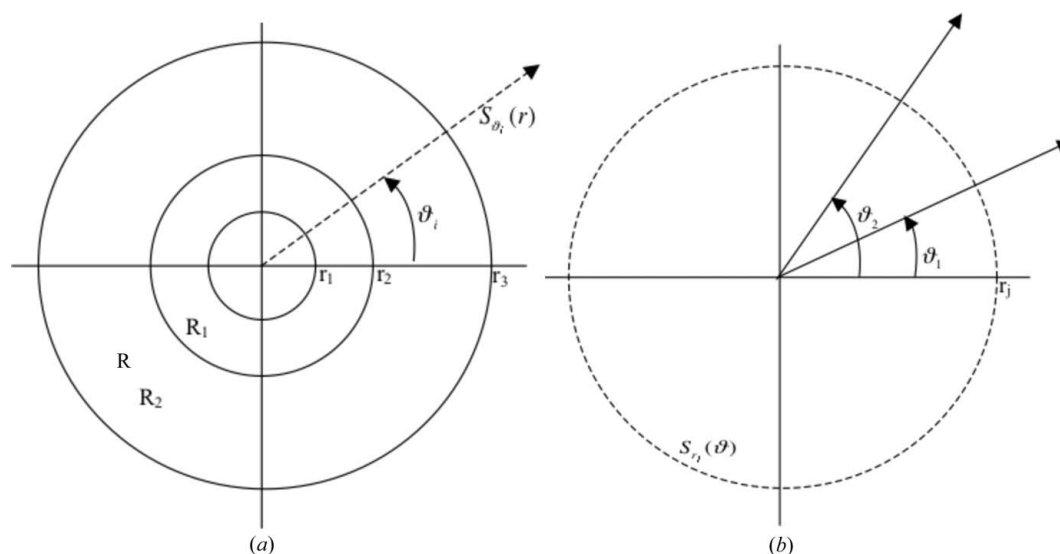


Figure 5

The two regions, R_1 and R_2 , over which $S_1(r)$ and $S_2(r)$ are modelled are shown in (a). The one-dimensional function, $S_{\theta_i}(r)$, is also shown (dotted line) for a fixed θ_i where $S_1(r) = \sum_{\theta_i=0}^{\pi} S_{\theta_i}(r)$ for $r_1 < r < r_2$, and $S_2(r) = \sum_{\theta_i=0}^{\pi} S_{\theta_i}(r)$ for $r_2 < r < r_3$. The wedge between two values of θ , θ_1 and θ_2 , is shown in (b). Here the one-dimensional function, $S_{r_j}(\theta)$, is shown (dotted line) for a fixed r_j , where $S(\theta) = \sum_{r_j=r_{\min}}^{r_{\max}} S_{r_j}(\theta)$.

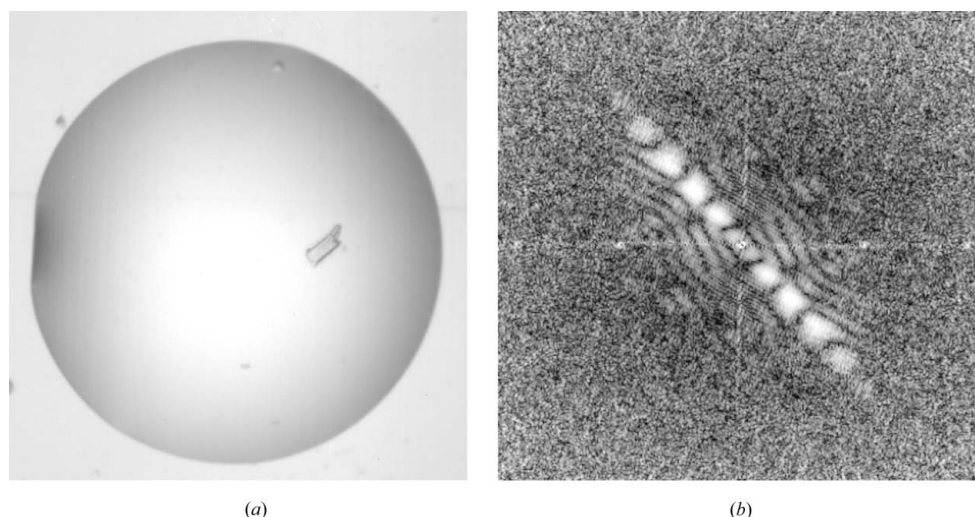


Figure 6

The FT of the image in (a) is shown in (b). The striped spoke of high intensity radiating out from the zero spatial frequency point is due to the single crystal in the drop.

(making the spectrum approximately flat) and the log-contrast, c_l , calculated by summing the logarithm of the power spectrum. Hence, these parameters are given by the equations:

$$c_w^2 = \sum_{(u,v)} r^2 S(u, v) \quad (7)$$

and

$$c_l^2 = \sum_{(u,v)} \log S(u, v). \quad (8)$$

The power spectra are modelled according to a power law and, for each radial segment the parameters estimated are: the

slope, α , of the power law; the intercept A of the power law; the r.m.s. contrast; the log contrast.

Following Papoulis (1965), the power spectrum, S , is separated into a product of a deterministic modulation function, M , and a probabilistic residue, ε (see Fig. 7), *i.e.*

$$S = M\varepsilon. \quad (9)$$

Although the ideal ε is 'probabilistic' in nature, deviations from the model vary and much information remains in ε . Thus statistical parameters provide useful descriptors and we also determine the following for each radial segment: the standard deviation of ε ; the skewness of ε ; the kurtosis of ε .

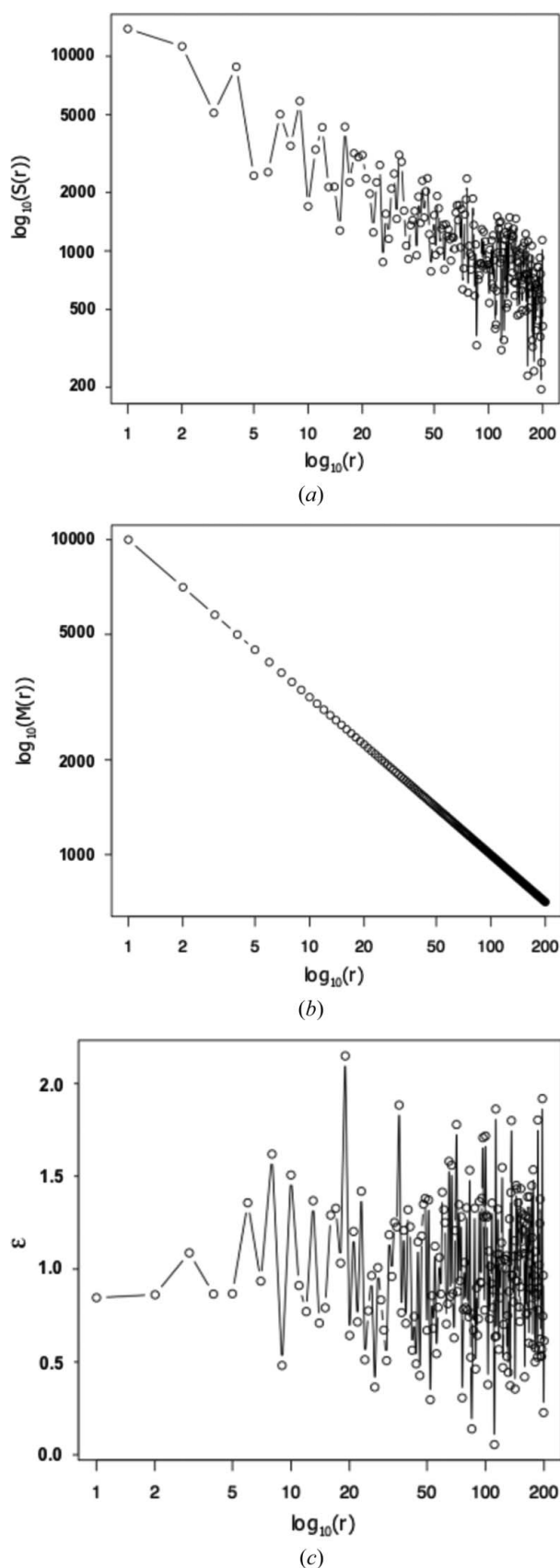


Figure 7
The log-log plot for a power spectrum, S , is shown in (a), with those corresponding to the fitted model, M , and residue, ϵ , in (b) and (c), respectively.

Hence, the total number of radial descriptors is 14 when two regions are used.

5.2. Angular descriptors

Distinct spikes in the function $S(\vartheta)$ at a range of radial distances can indicate the presence of crystals. In order to reduce the effect of the horizontal line in the plastic crystallization plate (see §3), any spokes in the vertical direction ($\pm 5^\circ$) are not considered and so, to determine angular descriptors, $S_{R_i}(\vartheta)$ is calculated for ϑ at every 1° from 5 to 85° , and from 95 to 175° , where

$$S_{R_i}(\vartheta) = \sum_{r=r_i}^{r_{i+1}} S(r, \vartheta). \quad (10)$$

The $S_{R_i}(\vartheta)$ are then multiplied together to obtain

$$S_T(\vartheta) = \prod_{i=1}^N S_{R_i}(\vartheta) \quad (11)$$

so that large values of the function $S_T(\vartheta)$ indicate angles, ϑ , where high intensities occur at both low and high spatial frequency and can therefore be used to detect spokes. The number of peaks in $S_T(\vartheta)$ above a pre-set threshold, ρ , and the intensity of the highest peak are both used as descriptors for classification.

A measure of the variation in intensity with ϑ is given by the standard deviation of the parameters A and α , fitted for each ϑ in the range above. High skewness suggests a few deviant peaks rather than many, smaller fluctuations and provides further information for classification. Other statistical measures have been investigated, but no improvement was achieved by adding higher moments such as kurtosis. Hence the descriptors calculated for each ring, R_i , are: the standard deviation of A ; the standard deviation of α ; the skewness of A ; the skewness of α . In addition we calculate: the number of peaks in $S_T(\vartheta) > \rho$; the intensity of the highest peak in $S_T(\vartheta)$.

Hence the number of angular descriptors is four times the number of rings plus two. In fact, for speed and simplicity, we have used only two rings as for the calculation of the radial descriptors giving just ten further descriptors.

6. Results and discussion

As crystallization experiments produce a continuum of results rather than discrete outcomes, the number of classes used is a somewhat arbitrary choice. The identification of crystals is the main aim, but a two-class system consisting only of crystals and non-crystals would give no information for subsequent trials. In the absence of crystals, other phenomena, such as crystalline precipitate, can indicate that conditions are close to those required for crystallization (see the website of Terese Bergfors: <http://alpha2.bmc.uu.se/terese/crystallization/library.html>). Such conditions can be refined in optimization protocols to produce diffraction-quality crystals and should therefore be identified. Equally, experiments resulting in heavy amorphous precipitate or denatured protein show that the

Table 1

Examples of experimental results associated with the seven categories used to classify crystallization images.

Class	Experimental result
0	Empty drop
1	Denatured protein
2	Amorphous precipitate
3	Oil drops; phase separation; crystalline precipitate
4	Micro-crystals
5	Crystal clusters
6	Single crystals

conditions are inappropriate and this information is also valuable for the development of automated screening procedures *via* data-mining techniques.

Studies to assess the reliability of human classification of images from crystallization experiments showed that agreement rates are surprisingly low and that the agreement rate between any two crystallographers decreases as the number of classes increases (Wilson & Berry, 2007). We found that a seven-class system gave an average agreement rate of ~70% for 16 crystallographers and that, in most cases where the scores do not agree, they differ by only one class. Allowing 'agreement' to tolerate a one-class difference gave an average agreement rate of ~94% and allowing a two-class difference gave close to total agreement at ~99%.

The descriptors described in the previous section were calculated for a training set of 700 images, classified by eye into the seven categories shown in Table 1, and used to train a learning vector quantization (LVQ) neural network (Kohonen *et al.*, 1992). The resulting LVQ code vectors were then used to classify an independent test set consisting of 150 images from each of the seven classes. The results are given in Table 2. The truth table showing the classification obtained using the Fourier descriptors in Table 2(a) can be compared with that in Table 2(b), which shows the results when the images are analysed using the object-based method of Wilson (2004). The diagonal entries give the percentage of exact matches with the scores given when the images were classified by eye. It can be seen that these results are generally higher for the FT method with the percentage of images correctly classified as containing precipitate in particular being much higher. However, misclassifications between neighbouring classes need not necessarily be considered incorrect if the continuous nature of the experiment and the diversity in human classification is taken into account and the bold values in the tables indicate acceptable classifications. More serious misclassifications appear further away from the diagonal with images classified too low above the diagonal and images classified too high below the diagonal. The use of parameters derived from frequency space leads to a higher proportion of false negatives in terms of crystals or crystal clusters being underrated as compared with the object-based method. For example, 8.7% of images containing good crystals are classed as uninteresting and another 5.3% as empty. This is not unexpected as the majority of the crystallization drop can be empty in the case of single crystals, particularly those growing at the edge of the

Table 2

Agreement rates between the scores given by *ALICE* and the scores given when classified by eye.

The columns correspond to the results of automated analysis for each class (according to the visual score) in the rows of the table so that the diagonal entries show exact matches. The figures given are percentages of the 150 images in each class (so that the values in each row sum to 100%) and bold values indicate acceptable classifications. The results in (a) were obtained using the Fourier descriptors for the crystallization drop and the results in (b) were obtained using classification variables calculated for individual objects within the drop.

	6	5	4	3	2	1	0
(a)							
Class 6	44.0	15.3	6.7	18.7	1.3	8.7	5.3
Class 5	18.0	39.3	12.0	14.7	2.0	9.3	4.7
Class 4	10.0	18.7	8.7	25.3	19.3	10.7	7.3
Class 3	4.0	0.7	3.3	42.7	40.0	6.7	2.7
Class 2	0.0	0.0	0.0	0.0	95.3	3.4	1.4
Class 1	4.7	6.1	4.7	23.6	6.1	54.7	0.0
Class 0	0.0	2.7	0.0	0.7	10.7	0.0	86.0
(b)							
Class 6	26.7	46.0	14.7	4.7	3.3	1.3	3.3
Class 5	7.3	47.3	18.7	9.3	10.7	4.7	2.0
Class 4	4.0	24.0	22.7	31.3	14.7	2.0	1.3
Class 3	1.3	6.0	10.0	28.0	47.3	3.3	4.0
Class 2	0.0	6.1	0.7	27.0	66.2	0.0	0.0
Class 1	0.7	6.1	7.4	22.3	18.2	44.6	0.7
Class 0	0.0	1.3	1.3	1.3	3.3	0.7	92.0

drop. In such cases, classification based on individual objects is likely to perform better. However, many images have features that occur throughout the drop and are better suited to Fourier methods. Fig. 8 shows an image of a drop containing precipitate together with its Fourier transform and the log-log plot showing the change in intensity with spatial frequency. In most cases, the relationship between spectral power and spatial frequency results in a relatively straight line in the log-log plot, but this is less true of precipitate images and, in fact, a low slope at low frequency and a high slope at high frequency gives a good indication of the presence of precipitate. Dividing the power spectra into two rings allows this feature to be exploited and results in high classification rates for precipitate images. It could be that more rings would provide further discrimination between classes, but the use of many more rings would result in many parameters and further investigation is necessary to determine the optimum number.

As the two methods exploit complementary features within the images, their combination leads to improved classification rates as shown in Tables 3 and 4. However, the best way to combine the results from the two methods is not obvious. Both methods provide a score between 0 and 6 for the image, with the image score for the object-based method obtained by combining the results from individual objects. A simple averaging of the two scores does give fewer very wrong classifications, such as crystals classed as class 3 or less, but also seems to spread the classification across more classes. Taking the maximum class obtained from either method dramatically improves the classification rate for crystals, with ~82% of all images containing crystals or crystal clusters (classes 5 and 6) being classified as either crystal or crystal cluster. However,

the number of false positives is also increased, with more empty and uninteresting drops being classed too high.

Rather than determining the class for each object or image from a single winning vector, the LVQ can be used to provide a probability for every class, calculated from the classes of the top ten vectors weighted according to the distance measure. The probability of each class according to the Fourier method

can then be used to multiply the object probabilities before the individual objects are evaluated. Each object is then assigned to the class with the highest probability and an overall image score obtained, taking into account the number of objects and the number of pixels in each class. The results are shown in Table 3(c). Although the method gives fewer false positives, the number of false negatives is also increased and, overall, the maximum class gives the best results.

Table 4 summarizes the results with the automated scores pooled to give just three output classes: those images that should definitely be looked at by eye, *i.e.* any images identified as crystalline; those images with precipitate or some slightly better outcome that should be looked at if no higher scores are obtained or if time permits; and those images which are most unlikely to show good results and need not be considered by the crystallographer.

It is clear from Table 4 that class 1 images, *i.e.* uninteresting or poor outcomes in terms of the crystallization experiment, are often the most difficult to classify correctly. As well as denatured protein that can take on various morphologies, images in this class can include foreign bodies, such as dust and fibres, bubbles and skin that can form over the drop and 'wrinkle' as evaporation takes place. Such images very often look more interesting in terms of the classification variables than they are in reality and are therefore classified too high. On the other hand, class 4 images, micro-crystals, are often classified as class 3, reflecting the situation found in human classification where these two classes caused most problems. However, we have found that using fewer classes merely moves the fuzzy boundaries and reduces the ability to sort the images, which is the main aim of this research.

The combination of the drop-based approach, *i.e.* the Fourier method, with the object-based method gives improved results, but the best method to combine the results needs further investigation. Using the maximum score from either method currently gives the best results, although, unsurprisingly, the number of false positives increases. However, the most important criteria for an automated scoring system must be low rates of false negatives, in particular missed crystals, even if this is at the expense of higher false-positive rates. The Fourier method is more accurate for images that are consistent throughout the drop, such as those containing precipitate and empty drops, but suffers from high rates of false negatives. Changing the class probabilities for individual objects using probabilities obtained for the whole image gives results that are dominated by the Fourier approach as every object's class is affected. A better way to combine the probabilities could, in principle, be devised and would ideally also allow the class probabilities of one object to influence those of another.

The aim is not to replace human classification, but to reduce the number of experiments that need to be examined by eye. Automated scoring allows the images to be examined in order of merit and, when high-scoring conditions are confirmed, no further images need be considered. The potential for the development of optimization procedures is also important; not only will many promising initial conditions be identified, but failed experiments will also be recorded, allowing the possi-

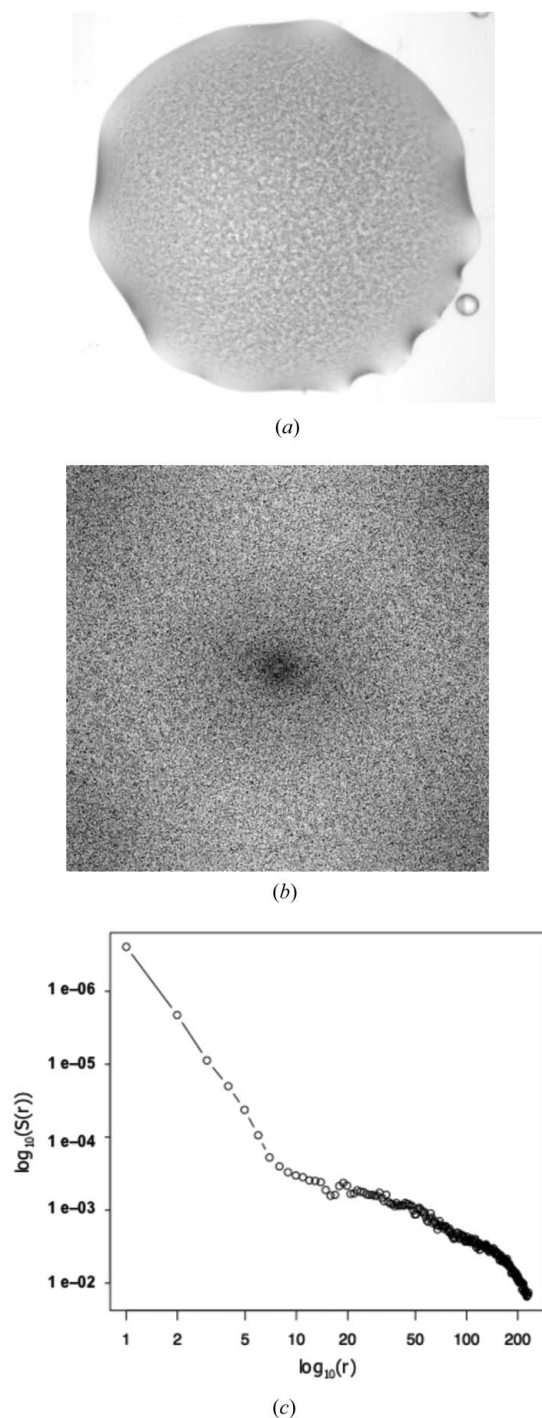


Figure 8

The image of a drop containing precipitate is shown in (a) with its FT in (b), and the log-log plot showing the change in intensity, $S(r)$, with spatial frequency, r , in (c).

Table 3

Classification rates for the scores given by *ALICE* when the object-based approach and the FT method are combined.

The columns correspond to the results of automated analysis for each class (according to the visual score) in the rows of the table and are given as a percentage of the 150 images in that class. Again, bold values indicate acceptable classifications. The results in (a) were obtained by averaging the scores from the two methods applied separately. The results in (b) were obtained by taking the maximum score from either method as the final score for the image. The results in (c) were obtained by multiplying the class probabilities for individual objects by the class probabilities from the FT method before the image score was determined in the object-based approach.

	6	5	4	3	2	1	0
(a)							
Class 6	39.3	22.0	22.0	12.7	1.3	1.3	1.3
Class 5	12.7	38.0	29.3	13.3	2.7	2.0	2.0
Class 4	4.7	20.0	26.0	32.0	14.7	1.3	1.3
Class 3	0.7	2.0	14.7	44.7	35.3	0.7	2.0
Class 2	0.0	0.0	6.1	26.4	66.9	0.7	0.0
Class 1	1.4	2.7	10.1	20.3	41.9	23.6	0.0
Class 0	0.0	0.0	0.7	3.3	4.7	10.7	80.7
(b)							
Class 6	56.0	29.3	8.0	4.0	0.7	0.7	1.3
Class 5	20.7	57.3	12.0	5.3	0.7	2.0	2.0
Class 4	12.0	30.7	22.0	26.7	6.7	0.7	1.3
Class 3	5.3	6.0	12.0	46.0	28.7	0.0	2.0
Class 2	0.0	6.1	0.7	27.0	66.2	0.0	0.0
Class 1	5.4	10.1	10.1	38.5	12.2	23.6	0.0
Class 0	0.0	4.0	1.3	2.0	12.0	0.0	80.7
(c)							
Class 6	40.7	26.7	10.7	8.7	6.0	3.3	4.0
Class 5	12.7	47.3	16.0	12.7	5.3	3.3	2.7
Class 4	8.0	22.0	14.7	26.7	20.0	4.7	4.0
Class 3	0.7	6.7	6.7	39.3	39.3	7.3	3.3
Class 2	0.0	0.0	0.0	2.0	97.3	0.7	0.0
Class 1	0.7	6.1	6.1	20.3	16.2	50.7	0.0
Class 0	0.0	1.3	0.0	0.7	5.3	0.0	92.7

Table 4

Classification rates for the scores given by *ALICE* when the object-based approach and the FT method are combined.

Here the results in Table 3 have been pooled and the three columns correspond to automated scores, indicating images that should definitely be looked at by the crystallographer (yes), those that perhaps should be looked at (maybe) and images that need not be looked at (no). Each row of the table shows the class according to the visual score. Again (a) shows results obtained by averaging the scores from the two methods applied separately, (b) shows results obtained by taking the maximum score from either method as the final score for the image and (c) shows results obtained by multiplying the class probabilities for individual objects by the class probabilities from the FT method before the image score was determined in the object-based approach.

	Yes	Maybe	No
(a)			
Class 6	83.3	14.0	2.6
Class 5	80.0	16.0	4.0
Class 4	50.7	46.7	2.6
Class 3	17.4	80.0	2.7
Class 2	6.1	93.3	0.7
Class 1	14.2	62.2	23.6
Class 0	0.7	8.0	91.4
(b)			
Class 6	93.3	4.7	2.0
Class 5	90.0	6.0	4.0
Class 4	64.7	33.4	2.0
Class 3	23.3	74.7	2.0
Class 2	6.8	93.2	0.0
Class 1	25.6	50.7	23.6
Class 0	5.3	14.0	80.7
(c)			
Class 6	78.1	14.7	7.3
Class 5	76.0	18.0	6.0
Class 4	44.7	46.7	8.7
Class 3	14.1	78.6	10.6
Class 2	0.0	99.3	0.7
Class 1	12.9	36.2	50.7
Class 0	1.3	6.0	92.7

bility for automated screening protocols to be developed *via* data-mining techniques.

Funding for CGW was provided by BIOXHIT [Biocrystallography (X) on a Highly Integrated Technology Platform for European Structural Genomics] under the 6th Framework Programme of the European Commission (LSHG-CT-2003-503420). Funding for JF was provided by CCP4 (Crystallographic Computing Project, Number 4). JW is a Royal Society University Research Fellow. The authors would like to thank D. Watts for useful comments.

References

Bern, M., Goldberg, D., Stevens, R. C. & Kuhn, P. (2004). *J. Appl. Cryst.* **37**, 279–287.
 Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C., Collins, R., Luft, J., DeTitta, G. & Jurisica, I. (2003). *Acta Cryst.* **D59**, 1619–1627.
 Delibasis, K., Undrill, P. E. & Cameron, G. G. (1997). *Signal Process.* **57**, 19–33. (See also <http://www.biomed.abdn.ac.uk/Abstracts/A00033/>.)
 Gonzalez, R. C. & Woods, R. E. (2002). *Digital Image Processing*, 2nd ed. New Jersey: Prentice Hall.
 Jurisica, I., Wolfley, J. R., Rogers, P., Bianca, M. A., Glasgow, J. I., Weeks, D. R., Fortier, S., DeTitta, G. T. & Luft, J. R. (2001). *IBM Syst. J.* **40**, 394–409.

Kohonen, T., Kangas, J., Laaksonen, J. & Torkkola, J. (1992). *Neural Netw.* **1**, 725–730. [The *LVQ_PAK* program can be downloaded from <http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml>.]
 Liu, F. (1997). *Modelling Spatial and Temporal Textures*, PhD thesis, Massachusetts Institute of Technology.
 Mayo, C. J., Diprose, J. M., Walter, T. S., Berry, I. M., Wilson, J., Owens, R. J., Jones, E. Y., Harlos, K., Stuart, D. I. & Esnouf, R. M. (2005). *Structure*, **13**, 175–182.
 Page, R. & Stevens, R. C. (2004). *Methods*, **34**, 373–389.
 Pan, S., Shavit, G., Penas-Centeno, M., Dong-Hui, X., Shapiro, L., Ladner, R., Riskin, E., Hol, W. & Meldrum, D. (2006). *Acta Cryst.* **D62**, 271–279.
 Papoulis, A. (1965). *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill.
 Saitoh, K., Kawabata, K., Asama, H., Mishima, T., Sugahara, M. & Miyano, M. (2005). *Acta Cryst.* **D61**, 873–880.
 Schaaf, A. van der & van Hateren, J. H. (1996). *Vis. Res.* **36**, 2759–2770.
 Spraggon, G., Lesley, S. A., Kreusch, A. & Priestle, J. P. (2002). *Acta Cryst.* **D58**, 1915–1923.
 Wilson, J. (2002). *Acta Cryst.* **D58**, 1907–1914.
 Wilson, J. (2004). *Crystallogr. Rev.* **10**, 73–84.
 Wilson, J. & Berry, I. M. (2005). *J. Appl. Cryst.* **38**, 493–500.
 Wilson, J. & Berry, I. M. (2007). In preparation.